MPEG-7: Formextraktion und -deskription

Daniel Steinhöfer
steinhoe@ifi.lmu.de

Universität München
Amalienstrasse 17, 80333 Munich, Germany

Zusammenfassung Mit einer stetig zunehmenden Masse an visuellen Mediendaten steigt der Bedarf, diese effizient jenseits von rein textueller Information und v.a. vollautomatisch zu kategorisieren und zu beschreiben. Zu diesem Zweck hat die MPEG den MPEG-7 Standard vorgeschlagen. Diese Arbeit stellt die vier im MPEG-7 Visual Standard zur Beschreibung von Formen innerhalb von Mediendaten festgehaltenen Deskriptoren zusammen. Die in der Ebene definierten Deskriptoren - der *Region-Based* und der *Contour-Based Shape Descriptor* basieren auf einer Contour Scale Space - Berechnung, bzw. auf einer Angular Radial Transformation. Der *2-D/3-D Shape Descriptor* dient der Beschreibung dreidimensionaler Objekte mit Hilfe von zweidimensionalen Schnappschüssen aus mehreren Blickwinkeln, welche mit einem - in der Ebene definierten - Deskriptor festgehalten werden. Der *3-D Shape Spectrum Descriptor* basiert auf dem Shape-Index um dreidimensionale Objekte statistisch angenähert zu beschreiben.
Beispielanfragen innerhalb der MPEG-7 Testdatenbank werden zur Veranschaulichung der Deskriptoren angeführt, sowie deren Güte in den deskriptorspezifischen relevanten Kategorien aufgezeigt. Abschließend wird ein praxisrelevantes Anwendungsbeispiel - eine webbasierte Videosuche - erläutert.

1 Einleitung

In Folge der weiten Verbreitung digitaler Aufzeichnungsgeräte wie Fotohandys, Digitalkameras oder Scannern, sowie der nach wie vor stetig zunehmenden Vernetzung von Rechnern mit immer höheren Bandbreiten, welche längst die Möglichkeit des komfortablen Austauschs von beliebigen Mediendaten bieten, steigt auch der Bedarf, auf Letztere effizient zugreifen zu können. Jedoch setzen Such- und Vergleichsanfragen eine Kategorisierung bzw. eine Beschreibung der Daten voraus.
Rein textuelle Beschreibungen, die den eigentlichen Mediendaten zugeordnet werden, sind die hier wohl gängigste Methode. Herkömmlicherweise basiert die Suche dann auf Textanfragen, mittels derer Mediendaten etwa in Datenbanken, Netzverzeichnissen oder auf dem heimischen Rechner aufgefunden werden sollen.
Mit steigender Datenmenge wird die vom Menschen zu tätigende textbasierte Beschreibung zunehmend arbeitsaufwendig und ineffizient. Zudem ist oftmals die Beschreibung und Kategorisierung nicht eindeutig durchführbar und erhält somit auch eine subjektive und damit unpräzise Komponente. Auch ist es in manchen Anwendungsfällen, z.B. in einer inhaltlich abstrakten Warenzeichendatenbank, schwer bis unmöglich, überhaupt textuelle Beschreibungen zu erstellen. Verallgemeinert können v.a. Bilddaten sehr inhaltsreich sein und textuelle Annotationen dem Inhalt nicht genügen.

Der Mensch hat die Fähigkeit, Gegenstände allein aufgrund ihrer Form zu erkennen und einzuordnen. Formen und Umrisse tragen also für den Menschen entscheidend-semantische Informationen, die textuell meist nur schwer zu erfassen sind. Automatische Extraktionsalgorithmen und kompakte Darstellungsweisen ermöglichen es, die Form als ein Kriterium zur Beschreibung von Inhalten von Mediendaten zu benutzen. Viele Anwendungen sind geradezu prädestiniert, auf Formen basierende Anfragen zu verwenden. Dabei kann die Information der Form zwei- oder dreidimensional vorliegen. Prinzipiell kann hierbei die Analyse zweidimensionaler Formen in die Kategorien umrissbasiert und regionsbasiert unterschieden werden. In der Ersteren werden die Umrisse eines Objekts betrachtet, in der Letzteren die Pixelverteilung innerhalb des gesamten Objekts.

In diesem Rahmen werden im MPEG-7 Visual Standard vier *Shape Descriptors* (SDs) definiert. Dabei decken die SDs die Anwendungsbereiche in der Ebene, sowie im Dreidimensionalen ab. Der 2-D/3-D SD dient der Beschreibung dreidimensionaler Objekte mit Hilfe von zweidimensionalen SDs aus mehreren Blickwinkeln. Der 3-D SD ist auf dreidimensionalen Gitternetzmodellen von Objekten definiert.

Im Folgenden werden zuerst bisherige Arbeiten auf dem Gebiet vorgestellt. Danach werden die vier SDs des MPEG-7 Visual Standards beschrieben. Abschließend wird eine praxisrelevante Anwendung - eine webbasierte Videosuche - vorgestellt.

2 Verwandte Arbeiten

Es gibt keine einheitliche Definition von *Form*, welche eine verbreitete und intuitive Bedeutung darstellt, die geometrische, sowie topologische und thematische Eigenschaften abdeckt. Die Form wird in diversen Theorien über die optische Wahrnehmung betrachtet, welche darauf zielen zu verstehen, wie der Mensch visuelle Formen aufnimmt. Der Psychobiologe D.O. Hebb [1] behauptet, dass der Mensch eine Form nicht als Ganzes, sondern als Zusammenhang aus Teilformen wahrnimmt, sowie dass die räumlichen Beziehungen zwischen den Teilformen eines Objekts zuerst erlernt werden müssen, um das Objekt erfolgreich erkennen zu können. Näher an visuellen Anwendungen am Rechner ist der hierarchische Ansatz, die Entwicklung von Formen mit sich ändernder Auflösung zu betrachten, welcher in [2] vorgeschlagen wird. Derartige Theorien stellen eine Basis für nützliche Ansätze dar, auf denen Methoden der rechnergestützten Formerkennung aufbauen können [3].
Formbasierte Objekterkennung wird als ein interessantes Problem in diversen Arbeiten aufgegriffen, u.a. in [4][5][6]. Die Herausforderung besteht darin, die Form auf eine robuste und möglichst kompakte Art und Weise zu beschreiben. Die folgenden Abschnitte beschreiben jeweils verwandte Arbeiten im Zweidimensionalen und Dreidimensionalen.

2.1 Zweidimensionale SDs

Das MPEG-7 Experimentation Model - die Testumgebung, welche die Güte und Performanz aller Deskriptoren des MPEG-7 Standards auf den Prüfstand stellt - enthielt vormals den *Zernike Moment Descriptor*. Dieser wurde später aufgrund seiner schlechteren Suchgenauigkeit durch den hier vorgestellten Region-Based SD ersetzt [7]. Der Zernike SD überträgt - ganz ähnlich dem Region-Based SD - das Bild auf eine Menge von orthogonalen Basisfunktionen im Komplexen. Allerdings hat dieser SD eine inhärent höhere Gewichtung der radialen Komponente, was der der menschlichen Wahrnehmung eher entsprechenden Betonung der winkelbezogenen Eigenschaften entgegensteht [7].

2.2 Dreidimensionale SDs

Im Bereich der dreidimensionalen SDs gibt es diverse Ansätze, welche in vier Kategorien zu unterteilen sind [8]. Prinzipiell werden hier Objekte betrachtet, welche als Gitternetzmodelle mit Flächen, Kanten und Punkten beschrieben sind oder mit Hilfe von weiteren geometrischen Primitiven spezifiert sind.

Strukturbasierte Ansätze. Die Idee hierbei ist es, die Zerlegung von Objekten in Untereinheiten zu nutzen, welche gewisse Homogenitätskriterien zu vordefinierten Merkmalen der Form erfüllen. Somit erhält man die Grundbausteine eines Objekts, die in einer Datenstruktur gespeichert werden, welche u.a. Nachbarschaftsbeziehungen darstellen kann [8]. In [9] wird ein SD vorgeschlagen, welcher auf Graphen von maximalen Teilstücken der Oberflächen eines Objekts aufbaut. Eine Funktion, welche die Hauptkrümmungen des Objekts berechnet, bildet Letztere auf einen Form-Index ab [10]. Der Zusammenhang der Teilstücke wird schließlich mittels eines Adjazenzgraphen kodiert [8]. Um ein Ähnlichkeitsmaß zwischen zwei dreidimensionalen Objekten zu erhalten, wird eine Graphvergleichstechnik [11] genutzt.

Abweichungsbasierte Ansätze. Die abweichungsbasierten Ansätze definieren das Abstandsmass zweier Oberflächen mittels der Energie, welche aufgebracht werden müsste, um die Flächen aneinander anzupassen und wird mit Hilfe der dynamischen Gleichung des Kräftegleichgewichts zwischen inneren und äußeren Kräften bestimmt [12].

Transformationsbasierte Ansätze. Auf Transformationen basierende Ansätze nutzen Integraltransformationen für die Darstellung von Formen. Dazu werden Techniken wie *SAI* [13] oder *harmonic maps* [14] genutzt, um die Oberfläche des Gitternetzmodels auf ein sog. *attribute image* [8] im Zweidimensionalen abzubilden. Die hierfür notwendige starke geometrische Gleichmäßigkeit von Gitternetzmodellen ist allerdings meist nicht gegeben und muss durch rechenaufwendige Vorverarbeitungsschritte geschaffen werden [8].

Statistische Ansätze. Statistische SDs bestehen aus Momenten [15][16][17] oder Verteilungen von deterministischen [17][5] oder zufälligen [16] geometrischen Primitiven [8]. Auch bei der Verwendung der statistischen Momente für die Formbeschreibung wird meist ein Vorverarbeitungsschritt in Form einer Normalisierung der Objektgröße und -position benötigt. Auch der MPEG-7 *3-D Shape Spectrum Descriptor* (3-D SSD) [3] gehört zur Klasse dieser Ansätze.
Die Art von SDs besitzen eine hohe Anfälligkeit gegenüber verschiedenen topologischen Darstellungen von Objekten und es bedarf deshalb ebenfalls einer vorangehenden Normalisierung des Gitternetzmodells des Objekts. Dazu können Koordinatenverteilungen [5] oder Histogramme mit Längen, Winkeln, Flächen und Volumina, welche jeweils Mengen von Sekanten, Dreiecken und Tetraedern zugeordnet sind, eingesetzt werden [8]. Diese Daten erhält man durch zufallsbasiertes Abtasten des ursprünglichen Gitternetzmodells. Jedoch sind diese Techniken anfällig für die verschiedenen möglichen Triangulationen, die von ein und demselben Gitternetzmodell existieren können.
Für die oberflächenbasierten *Extended Gaussian Images* (EGI) [18] wird eine Funktion auf einer Einheitskugel definiert, welche Informationen wie die Orientierung der Normalen, die Fläche oder Krümmung aufbaut [8]. Diese Ansätze sind empfindlich bezüglich der Ausrichtung der Flächen. So ist es möglich, dass zwei Objekte mit ähnlichen globalen Eigenschaften, aber mit einigen verschieden-orientierten Oberflächen von zwei komplett unterschiedlichen EGIs beschrieben werden [5][8].
Alle diese SDs bieten den Vorteil einer sehr kompakten Darstellung, die mit einem relativ geringen Rechenaufwand berechnet werden kann. Sie genügen aber, bis auf den *3-D SSD*, im Allgemeinen nicht der Forderung invariant gegenüber geometrischen Variationen zu sein [8]. Deshalb werden hier normalerweise einige vorverarbeitende Schritte zur räumlichen Anordnung, basierend auf der *Principal Component Analysis* (PCA), welche Vektoren durch ihre entscheidenden Komponenten ausdrückt, oder globale Maße wie minimal umgebende Quader vorgeschlagen [8].

3 MPEG-7 Shape Descriptors

Die vom MPEG-7 Visual Standard beschriebenen SDs können für zweidimensionale Objekte, also Bilder in jedem Kontext, sowie für dreidimensionale Objekte verwendet werden. Die dreidimensionalen Objekte können entweder als Gitternetzmodell vorliegen oder auch eine Ansammlung von Bildern aus verschiedenen Blickwinkeln auf das zu beschreibende Objekt sein. Die meisten Objekte in unserer Welt sind dreidimensional, wohingegen Bilder und Video 2-D Projektionen auf die Ebene darstellen. Man erkennt hier schon die Notwendigkeit, dass die Formbeschreibung möglichst invariant gegenüber Skalierung, Rotation und Translation sein muss. Aber auch im Dreidimensionalen muss aufgrund der möglichen Diversität zweier Gitternetz-Darstellungen eines Objekts gewährleistet sein, dass ein SD invariant gegenüber geometrischen Verschiedenheiten topologisch gleicher Objekte ist.
Die Deskriptoren beinhalten im Gegensatz zu den meisten Anderen, im MPEG-7 Standard definierten, Deskriptoren auch die Extraktion der zu speichernden

Features. Wobei der Standard nicht beschreibt, wie die Formen bzw. die Objekte aus einem konkreten Mediendatum herauszulösen sind - beispielsweise wie das Teilbild eines Autos in einer komplexen Szenerie aus einem Bild zu separieren. In den folgenden Abschnitten werden die vier MPEG-7 SDs im Einzelnen beschrieben.

3.1 Region-Based Shape Descriptor

Der Region-Based Shape Descriptor wurde für die Beschreibung von Objekten im Zweidimensionalen entwickelt. Hierbei wird die Pixelverteilung einer Form bzw. eines Objekts in der Ebene betrachtet. Dazu werden innere, sowie Pixel, die auf dem Rand der Form liegen, berücksichtigt. Dadurch kann dieser SD auf Objekte angewandt werden, die aus einer zusammenhängenden Region bestehen, sowie auf solche, die sich aus mehreren unzusammenhängenden Regionen, welche evtl. Löcher aufweisen, zusammensetzen [19]. Abbildung 1 zeigt Objekte für deren Beschreibung sich der Region-Based SD eignet.

Abbildung 1. Objekte für deren Beschreibung sich der Region-Based SD eignet. ([20])

Funktionsweise. Der SD basiert auf der *Angular Radial Transform* (ART). Diese besteht aus orthonormalen, sinusförmigen Basisfunktionen in Polarkoordinaten im Komplexen. Die Form wird in die Basisfunktionen zerlegt und die Werte der Koeffizienten normalisiert und quantisiert [19].

Angular Radial Transform. Die Angular Radial Transform (ART) ist eine orthogonale Einheitstransformation, die auf einer Einheitskreisscheibe definiert wird. Im Folgenden ist F_{NM} ein ART Koeffizient der Ordnung n und m, $f(\rho,\theta)$ eine Bildfunktion in Polarkoordinaten.
Die ART Koeffizienten sind folgendermaßen definiert:

$$F_{nm} = \langle V_{nm}(\rho,\theta), f(\rho,\theta)\rangle = \int_0^{2\pi}\int_0^1 V_{nm}^*(\rho,\theta)\, f(\rho,\theta)\, \rho \, \mathrm{d}\rho \, \mathrm{d}\theta$$

$V_{nm}(\rho,\theta)$ ist die Basisfunktion von ART, welche sich in eine radiale und eine den Winkel betreffenden Funktion aufteilen läßt:

$$V_{nm}(\rho,\theta) = A_m(\theta)\, R_n(\rho)$$

Die radiale Funktion basiert auf einer Cosinus-Funktion:

$$R_n(\rho) = \begin{cases} 1 & n = 0 \\ 2\cos(\pi n \rho) & n \neq 0 \end{cases}$$

Die Werte der ART sind invariant gegenüber Rotationen. Das wird durch das Verwenden einer Exponentialfunktion für die Winkel-Basisfunktion erreicht und wird im Darauffolgenden gezeigt [7]:

$$A_{nm}(\theta) = \frac{1}{2\pi} \exp(jm\theta)$$

Sei $f^{\alpha}(\rho, \theta)$ das im Ursprung um den Winkel α gedrehte Bild von $f(\rho, \theta)$, also $f^{\alpha}(\rho, \theta) = f(\rho, \alpha + \theta)$.
Dann ist die ART des gedrehten Bildes folgendermaßen gegeben:

$$F_{nm}^{\alpha} = \frac{1}{2\pi} \int_0^{2\pi} \int_0^1 V_{nm}^*(\rho, \theta) f^{\alpha}(\rho, \theta) \rho \, \mathrm{d}\rho \, \mathrm{d}\theta$$

bzw.

$$F_{nm}^{\alpha} = F_{nm} \exp(-jm\alpha)$$

Also ist der Wert der ART des gedrehten, sowie des originalen Bild derselbe mit

$$\|F_{nm}^{\alpha}\| = \|F_{nm}\| .$$

Darstellung des SD. Der ART SD wird als Menge von normalisierten Werten der komplexen ART Koeffizienten definiert. Es werden zwölf winkelbezogene und drei radiale Funktionen benutzt. Um die Skalierung der Formen zu normalisieren teilt man die Koeffizienten durch den Wert des ART Koeffizienten für n = 0, m = 0. So wird letzterer Koeffizient nicht als Teil des Deskriptors verwendet, da dieser nach der Normalisierung konstant ist. Schließlich werden die Koeffizienten auf 4 Bit quantisiert, um den Deskriptor kompakt zu halten [19]. Zusammengefasst entsteht bei Anwendung des Region-Based SD auf ein Objekt k ein Array $M[k]$ mit den normalisierten, auf 4 Bit quantisierten Werten der 35 ART Koeffizienten (3*12 - 1 für n,m = 0). Der SD ist somit 140 Bit groß.

Ähnlichkeitsmaß. Den Abstand zweier Objekte bzw. Formen d und q wird durch Anwendung der L_1-Norm berechnet:

$$Abstand_{d,q} = \sum_{i=1}^{35} \|M_d[i] - M_q[i]\|$$

Experimentelle Ergebnisse. Der Region-Based SD liefert insgesamt gute Ergebnisse mit der MPEG-7 Test-Datenbank, wie in Tabelle 1 zu sehen ist. Vor allem in den Disziplinen der Rotation und Skalierung werden gute Suchergebnisse geliefert. Bei den perspektivischen Verzerrungen dagegen legt der SD ein etwas schlechteres Ergebnis zu Tage. In Anbetracht der Komplexität der Test-Datenbank ist das Ergebnis bei der ähnlichkeitsbasierten Suche über einen klassifizierten Datensatz ebenfalls als gut zu bewerten [19].

Tabelle 1. Experimentelle Ergebnisse des Region-Based SD in der Warenzeichen - Datenbank. (Nach [19])

Test - Datensatz	Digital skaliert	Digital rotiert	Sensorisch skal./rot.	Perspektive	Klassifiziert	Insgesamt*
	86,20%	98,06%	98,76%	62,65%	62,13%	76,28%

*Klassifiziert wird vierfach gewertet

3.2 Contour-Based Shape Descriptor

Der Contour-Based SD basiert auf der Analyse der Umrisse eines Objekts bzw. einer Form. Objekte bei denen sich die entscheidende Form-Information hauptsächlich im Umriß wiederspiegelt, werden von diesem SD effizient beschrieben. Objekte, die aus mehreren unzusammenhängenden Regionen bestehen, können regionsweise mit dem SD und mit Hilfe von *Description Schemes* [21] beschrieben werden.

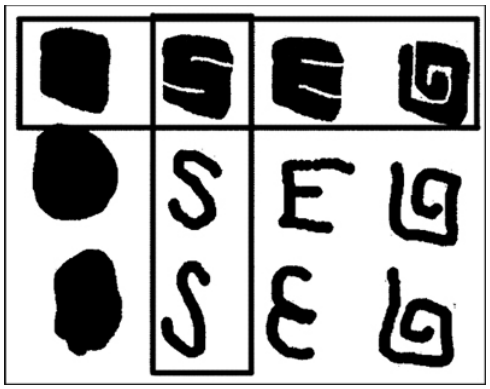

Abbildung 2. Vergleich von Contour- und Region-Based SD: Die Objekte der oberen Zeile - relativ flächig gefüllte Regionen - würden dem Region-Based SD nach als ähnlich gelten, während dagegen der Contour-Based SD diese als unähnlich werten würde und z.B. die S-Förmigen Objekte in der zweiten Spalte als ähnlich ansehen würde. ([20])

Funktionsweise. Der SD basiert auf der *Curvature Spale-Space* (CSS) Repräsentation des Umrisses [22][23]. Dieser wurden globale Parameter für die Form, sowie eine neue Quantisierung hinzugefügt. Zudem wurden die Feature-Vektoren in den entsprechenden Parameterraum übertragen, um die Performanz zu steigern [20].

CSS Repräsentation. Die CSS Repräsentation besteht aus dem Zerteilen der Umrisse einer Form in konvexe und konkave Bereiche, ähnlich wie der Mensch beim Vergleichen zweier Formen vorgeht [19]. Bei der CSS Repräsentation wird dies durch das Auffinden von Wendepunkten im Umriß realisiert. Die Analyse des Umrisses wird unter verschiedenen Abstufungen getätigt, welche durch

einen Glättungsprozeß entstehen. Vereinfacht werden hierfür zu Begin N auf dem Umriß im gleichen Abstand verteilte Punke zufällig gewählt. An diesen wird in jeder Iteration ein Tiefpassfilter angewandt, welcher die Kontur abflacht. So verschwinden die Wendepunkte nach und nach. Verschwindet ein Wendepunkt in einer Iteration dann entsteht ein sogenannter Peak. Der Vorgang wird solange wiederholt bis alle Wendepunkte verschwunden sind.
Mit einem *CSS Image* kann man den Sachverhalt bildlich darstellen: Die Wendepunkte werden *auf* der x-Achse aufgetragen und zwar gemäß ihrer Position auf dem Umriß im Uhrzeigersinn. Die Länge des Umrisses wurde vorher normalisiert. Nun wird für jede Iteration i die Position der Wendepunkte neu ermittelt. Die y-Koordinate entspricht i. Verschwindet ein Wendepunkt entsteht eine Spitze von zwei aufeinandertreffenden Wendepunkten. Abbildung 3 zeigt ein solches CSS Image mit 80 Iterationen und den entsprechenden Wendepunkten.

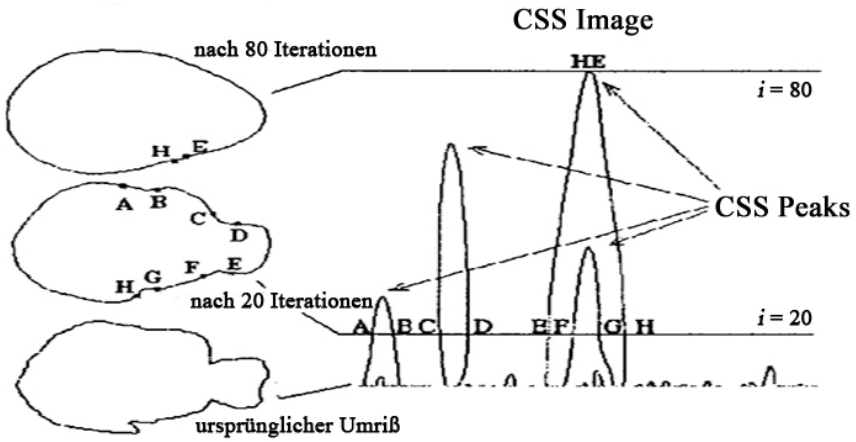

Abbildung 3. Beispiel für ein CSS Image mit 80 Iterationen und den Wendepunkten A - H. (Nach [20])

Darstellung des SD. Alle Felder des SD sind auf sechs bzw. sieben Bit quantisiert. Der SD besteht aus einem Feld für den höchsten Peak (7 Bit), einem Feld für die Anzahl der Peaks, einem Array für die x-Position der Peaks auf dem Umriß relativ zum höchsten Peak im Uhrzeigersinn, einem Array für die Höhe des Peaks relativ zum vorangegangen und, für jeweils den Ausgangsumriß und den Umriß nach der Berechnung des SD, den Werten für die Exzentrizität und die Rundheit des Umrisses.

Berechnung der Exzentrizität.

$$\text{Exzentrizität} = \sqrt{\frac{i_{20} + i_{02} + \sqrt{i_{20}^2 + i_{02}^2 - 2i_{20}i_{02} + 4i_{11}^2}}{i_{20} + i_{02} - \sqrt{i_{20}^2 + i_{02}^2 - 2i_{20}i_{02} + 4i_{11}^2}}}$$

mit $i_{02} = \sum M_{k=1}(y_k - y_c)^2$, $i_{11} = \sum M_{k=1}(x_k - x_c)(y_k - y_c)$, $i_{20} = \sum M_{k=1}(x_k - x_c)^2$, M ist die Anzahl der Punkte auf dem Umriß, (x_c, y_c) der Schwerpunkt des Objekts.

Berechnung der Rundheit.

$$\text{Rundheit} = \frac{\text{Umfang}^2}{\text{Fläche}}$$

Insgesamt hat der SD eine durchschnittliche Größe von 112 Bits pro Umriß.

Experimentelle Ergebnisse. Bei den Skalierungs- und Rotationtests erreicht der SD erwartungsgemäß gute Werte. Vor allem die 100% beim Test der Robustheit gegenüber Rotationen lassen sich dadurch erklären, dass die Peaks immer an derselben Stelle ausgehend vom höchsten Peak gespeichert werden. Die wahre Stärke des SD liegt allerdings bei der Ähnlichkeitssuche, bei denen der SD 79,15% in der MPEG-7 Testumgebung erreicht, wie auch in Tabelle 2 nachzuvollziehen ist [19].

Tabelle 2. Experimentelle Ergebnisse des Contour-Based SD in der MPEG-7 Testdatenbank. (Nach [19])

Test - Datensatz	Skalierung	Rotation	Ähnlichkeit	Nicht starr	Insgesamt
	91,03%	100%	79,15%	96%	90,22%

3.3 3-D Shape Descriptor

Die dritte Dimension am Rechner, als Modell der realen Welt, ist zweifelsfrei gerade im Zuge der rasanten Entwicklungen in den Bereichen der *Virtuellen Welten*, sowie der *Augmented Reality* eine wichtiger und interessanter Fokus im Rahmen des MPEG-7 Frameworks. Gänzlich im Dreidimensionalen arbeitet der 3-D SD. Dieser ist auf dreidimensionalen Gitternetzdarstellungen - wie etwa VRML - von Objekten definiert und basiert auf dem *Shape-Spectrum* Konzept, welches die im Folgenden beschriebene Erweiterung des *Shape-Index* darstellt.

Funktionsweise. Die Gitternetzdarstellung setzt sich aus einer Menge von Punkten mit Koordinaten im Dreidimensionalen, sowie einer Menge von Flächen zusammen, welche durch ein Array von Punkten beschrieben werden. Zusätzlich können Gitternetzmodelle auch Farb- oder Texturinformationen oder geometrische Informationen, z.B. die Normalen, enthalten.
Das 3-D Shape Spectrum besteht aus einem Histogramm, welches die Werte der Shape-Indizes des gesamten Gitternetzmodells beinhaltet.

Shape-Index. Der Shape-Index wurde ursprünglich als Maß für die lokale Konvexität von Flächen bei dreidimensionalen Formen in 3-D Gitternetzmodellen entwickelt [24][10]. Er basiert auf einer Funktion der beiden Hauptkrümmungen - der größten und kleinsten Krümmung - im Punkt p einer Fläche Σ. Dabei

bezeichnen k_p^1 und k_p^2 die Hauptkrümmungen im Punkt p. Der Shape-Index I_p im Punkt p wird folgendermaßen definiert:

$$I_p = \frac{1}{2} - \frac{1}{\pi} \arctan \frac{k_p^1 + k_p^2}{k_p^1 - k_p^2} \qquad \text{mit } k_p^1 \geq k_p^2$$

Der Shape-Index liegt im Intervall [0,1) und ist für plane Oberflächen nicht definiert. Er ist invariant gegenüber Skalierungen und Euklidischen Transformationen. Abbildung 4 zeigt fünf grundlegende geometrische Krümmungen mit den entsprechenden Werten des Shape-Index.

Bei Gitternetzmodellen von dreidimensionalen Objekten können diverse Irregu-

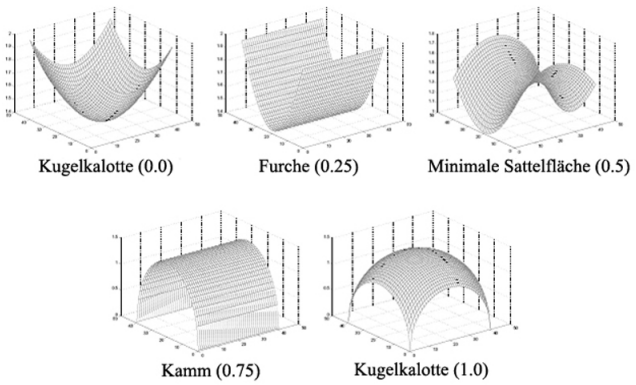

Abbildung 4. Grundlegende geometrische Krümmungen mit zugehörigem Shape-Index. (Nach [3])

laritäten auftreten, die in Vorverarbeitungsschritten ausgefiltert werden müssen. Zum einen spielt die Orientierung der Oberflächen eine entscheidende Rolle für geometrische Berechnungen. Diese ist durch die Umlaufrichtung der sie definierenden Punkte festgelegt. Ein Gitternetzmodell heißt orientierbar, wenn alle Flächen, welche sich mindestens eine Kante teilen, diese Kante(n) in verschiedenen Richtungen durchlaufen. Nur von orientierbaren Gitternetzmodellen können die Shape Indizes bzw. die Hauptkrümmungen berechnet werden.

Zum anderen muss die Anzahl der Zusammenhangskomponenten minimal gehalten werden, sowie degenerierte (Polygone ohne Fläche) und doppelt vorhandene Polygone entfernt werden. Dies wird durch eine regularisierende Filterung des ursprünglichen Gitternetzmodells erreicht. Um eine gleichmäßig-topologische Darstellung zu erreichen, werden die Modelle trianguliert, und anschließend die Verteilung und Größe der Dreiecke mit dem *Loop's subdivision*-Algorithmus [25] verfeinert und gleichmäßiger gestaltet. Dieser fügt auf den Kanten der Dreiecke des Modells mittig einen neuen Punkt ein und führt danach eine Tiefpassfilterung der Punktkoordinaten zur gleichmäßigeren Verteilung der Punkte durch.

Schätzung der Hauptkrümmungen. Die Schätzung der Hauptkrümmungen ist der Schlüsselschritt der 3D SD-Berechnung [3]. Die Hauptkrümmungen sind

durch die Eigenwerte der Weingarten-Abbildung definiert:

$$W = I^{-1}II$$

wobei I und II die erste bzw. zweite Fundamentalform darstellen. Letztere kann man durch eine polynomielle Oberflächenanpassung zweiten Grades abschätzen.

Berechnung der parametrischen Oberflächenanpassung. Dazu wird zuerst der Hauptnormalenvektor $\widetilde{N}_{f_i}$ einer Fläche f_i berechnet. Dieser stellt den gewichteten Durchschnitt der Normalenvektoren aller benachbarten Flächen dar.

$$\widetilde{N}_{f_i} = \frac{\sum_{f_k \in F_0\{f_i\}} w_k N_{f_k}}{\left\| \sum_{f_k \in F_0\{f_i\}} w_k N_{f_k} \right\|}$$

Wobei $F_0\{f_i\}$ die Menge der benachbarten Flächen von f_i und N_{f_k} den Normalenvektor der Fläche f_k bezeichnet. Die Gewichtung w_k wird entsprechend der Größe der Fläche f_k gewählt, um Ungleichmäßigkeiten in der Flächenverteilung des Gitternetzmodells auszugleichen. Für $\|.\|$ wird die L_2-Norm eines Vektors im R^3 verwendet.
Im Schwerpunkt der Fläche f_i wird der Nullpunkt eines in der z-Achse an dem Hauptnormalenvektor $\widetilde{N}_{f_i}$ ausgerichteten, kartesischen Koordinatensystems definiert. Die Zentroiden der Fläche f_i, sowie der benachbarten Flächen f_k werden in diesem Koordinatensystem als Menge $\{(x_i, y_i, z_i)\}_{i=1}^N$ erfasst.
Die parametrische Oberflächenanpassung wird durch das Angleichen einer quadratischen Oberfläche mit der Menge der Punkte $\{(x_i, y_i, z_i)\}_{i=1}^N$ erreicht [3].
Nun stellt $S = (x, y, z) = (x, y, F_a(x, y))$ die polynomielle Oberflächengleichung zweiten Grades dar. Mit

$$f_a(x, y) = a_0 x^2 + a_1 y^2 + a_2 xy + a_3 x + a_4 y + a_5 \ ,$$

$$a = (a_0 \ a_1 \ a_2 \ a_3 \ a_4 \ a_5)^t \text{ und}$$

$$b(x, y) = (x^2 \ y^2 \ xy \ x \ y \ 1)^t \text{ ergibt sich:}$$

$$f_a(x, y) = a^t b(x, y).$$

Der Parametervektor a wird durch die Anwendung der stochastischen *Methode der kleinsten Quadrate* bestimmt. Mit der Menge der Punkte der Zentroiden $\{(x_i, y_i, z_i)\}_{i=1}^N$ und den entsprechenden Gewichtungen $\{w_i\}_{i=1}^N$ entspricht der Parametervektor $\widehat{a}$ dem optimalen Wert für a:

$$\widehat{a} = \frac{\sum_{i=1}^N w_i z_i b(x_i, y_i)}{\sum_{i=1}^N w_i b(x_i, y_i) b^t(x_i, y_i)} = \arg\min_{a \in R^6} \sum_{i=1}^N w_i (z_i - f_a(x_i, y_i))^2$$

An dieser Stelle sieht man, dass der SD aufgrund der lokalen Koordinatensysteme pro Fläche invariant gegenüber Euklidschen Transformationen ist [3].
Schließlich können die Eigenwerte der Weingarten-Abbildung mit Hilfe der folgenden Gleichungen bestimmt werden:

$$I = \begin{pmatrix} 1 + a_3^2 & a_3 a_4 \\ a_3 a_4 & 1 + a_4^2 \end{pmatrix} \qquad , \qquad II = \frac{1}{\sqrt{1 + a_3^2 + a_4^2}} \begin{pmatrix} a_0 & a_2 \\ a_2 & a_1 \end{pmatrix}$$

Zusammenfassend werden die Hauptkrümmungen aller Punkte des Gitternetzmodells abgeschätzt und damit die Shape Indizes aller Punkte berechnet.

Darstellung des SD. Der SSD besteht aus dem Histogramm der Werte der Shape-Indizes für jeden Punkt im Gitternetzmodell. Dabei werden die Werte der Shape-Indizes auf n_{bins} quantisiert, also das Intervall [0, 1) des Wertebereichs des Shape-Index gleichmäßig in n_{bins} Intervalle eingeteilt:

$$\{\Delta_k\}_{k=1}^{N_{bins}} \quad \text{mit} \quad \Delta_k = \left[\frac{k-1}{N_{bins}}, \frac{k}{N_{bins}}\right)$$

$$\text{für} \quad k = \overline{1, N_{bins}-1} \quad \text{und} \quad \Delta_{N_{bins}} = \left[\frac{N_{bins}-1}{N_{bins}}, 1\right].$$

Die Größe jeder Fläche des Gitternetzmodells des Shape-Index, der zum Intervall Δ_k gehört, wird relativ zur Gesamtfläche des Gitternetzmodells zur k^{ten} Komponente des Histogramms hinzugefügt [3].
Der relative Flächenanteil der planaren Flächen, für welche der Shape-Index nicht definiert ist, sowie der relative Flächenanteil der Oberflächen, die weniger als N_{min} benachbarte Flächen haben, die ebenso nicht für die Berechnung eines Shape-Index geeignet sind, wird mit zwei separaten Histogrammbalken festgehalten. Planare Flächen werden mit Hilfe des Grads der Krümmung $C = AREA(F_0\{f_i\}) * k_a$ mit $k_a = \sqrt{k_1^2 + k_2^2}$ bestimmt, der eine Fläche als plan kennzeichnet, falls ein Schwellwert T unterschritten wird.
Insgesamt enthält der 3-D SD N_{bins} Klassen für das Shape Spectrum mit jeweils auf $BitsPerBin$ quantisierten Balken. N_{bins} ist mit acht Bit kodiert, $BitsPerBin$ mit vier Bit, wobei der maximale Wert bei für $BitsPerBin$ bei zwölf liegt. Dazu kommen die zwei Balken für die planaren und die alleinstehenden Oberflächen.

Ähnlichkeitsmaß. Die Ähnlichkeit zweier Objekte in Gitternetzdarstellung erfolgt durch Berechnung der L_1-Norm oder der L_2-Norm über alle Histogrammbalken der beiden 3-D SD. Die planaren, sowie die alleinstehenden Flächen können dabei eventuell außer acht gelassen werden.

Experimentelle Ergebnisse. Der 3-D SD wurde mit Hilfe eines Datensatzes bestehend aus 1290 3-D Gitternetzmodellen bewertet. Aus diesem wurden 228 Modelle ausgewählt und in 15 Kategorieren eingeteilt. Die Objekte wurden alle manuell vorverarbeitet. Sie wurden visuell inspiziert und dabei falsch orientierte Flächen korrigiert. Die L_1-Norm wurde zur Berechnung der Ähnlichkeit benutzt, wobei die Histogrammbalken für die planaren und die alleinstehenden Flächen nicht mit ausgewertet worden sind. Des Weiteren wurde das Shape Spectrum-Histogramm normalisiert, so dass $\sum_{i=1}^{N_{bins}} SSD(i) = 1$. Für jede der 15 Kategorien wurde die durchschnittliche Bull-Eye-Percentage (BEP) Punktzahl berechnet, indem eine Anfrage für jedes Objekt einer Kategorie gestartet worden ist. Q bezeichnet im Folgenden die Anzahl der vorklassifizierten Objekte einer Kategorie. Dann berechnet sich die BEP Punktzahl pro Anfrageobjekt aus der Anzahl der zur selben Kategorie gehörenden Objekte unter den ersten $2Q$ Treffern. Die durchschnittliche BEP Punktzahl ist 85%, was für die guten Unterscheidungskriterien des 3-D SD spricht. Auf Eigenschaften basierende

Tabelle 3. BEP Punktzahlen innerhalb der einzelnen Kategorien. Q entspricht der Anzahl der Objekte innerhalb einer Kategorie. (Nach [3])

Kategorie	**Q**	**BEP (%)**	**Kategorie**	**Q**	**BEP (%)**
4-Gliedrige	31	73	E1_Mx	9	100
Autos	17	66	Finger	30	100
Aerodynamisch	36	71	Buchstabe A	10	90
Bäume	21	56	Buchstabe B	10	100
Raketen	10	87	Buchstabe C	10	100
Ballone	7	95	Buchstabe D	10	100
Gebäude	10	39	Buchstabe E	10	100
Soma	7	100			

Kategorien wie die der modellierten Buchstaben schneiden dabei exzellent mit BEP Punktzahlen über 90% ab. Kategorien, die Eigenschaften mit semantischen Konzepten kombinieren - wie etwa 'Raketen' oder '4-Gliedrige' - erzielen immer noch BEP Werte zwischen 70% und 90%. In den Kategorien, die fast nur auf semantischen Konzepten aufbauen - wie 'Bäume' - erreicht der SD lediglich Werte im Bereich von 56% oder darunter [3]. Tabelle 3 veranschaulicht den vorangegangenen Sachverhalt.
Um die Größe und die Komplexität des 3-D SD zu vermindern wurde untersucht wie eine zunehmende Quantisierung, sowie die Anzahl der verwenden Histogrammbalken sich auf die globale durchschnittliche BEP Punktzahl auswirkt. In Tabelle 4 sieht man, daß bei einer Quantisierung auf neun Bit die globale BEP Punktzahl im Gegensatz zu einem in Gleitkommagenauigkeit beschriebenen 3-D SD nur um ein Prozent auf 84% sinkt und diese selbst bei sieben Bit pro Histogrammbalken auf dem relativ hohen Wert von 82% bleibt. Bei einer

Tabelle 4. Einfluß der Quantisierung in Bit auf die globale BEP Punktzahl. (Nach [3])

	FPP	b=12	b=11	b=10	b=9	b=8	b=7
Globale BEP	85	84	84	84	84	83	82

gleichbleibenden Quantisierung von zwölf Bit pro Histogrammbalken und einer gleichzeitigen Verminderung der Anzahl an Histogrammbalken von 100 auf 50 bleibt die globale BEP Punktzahl bei 84%. Erst wenn die Anzahl der Balken auf 25 bzw. zehn reduziert wird, sinkt die globale BEP Punktzahl auf 83% respektive 80% (siehe Tabelle 5).

Tabelle 5. Einfluß der Anzahl der Histogrammbalken bei gleichbleibender Quantisierung von zwölf Bit pro Balken auf die globale BEP Punktzahl. (Nach [3])

N_{bins}	100	50	25	10
Globale BEP	84	84	83	80

3.4 2-D/3-D Shape Descriptor

Dieser SD wurde entworfen, um dreidimensionale Objekte zu beschreiben. Dazu werden diese aus mehreren zweidimensionalen Ansichten festgehalten, die wiederum mittels eines SDs, der auf zweidimensionalen Ansichten arbeitet, beschrieben werden. Es können auch Farb- oder Texturdeskriptoren des MPEG-7 Visual Standards verwendet werden. Die Ähnlichkeit zweier Objekte ergibt sich dann aus dem Vergleich von Paaren von Ansichten der beiden Objekte mittels des Ähnlichkeitsmaßes des verwendeten SDs. Gute Ergebnisse wurden in diesem Zusammenhang mit dem Contour-Based SD erreicht [19].

4 Anwendungsbeispiel - Webbasierte Video-Suche

Exemplarisch für alle SDs wird nun ein Anwendungsbeispiel für den Contour-Based SD vorgestellt.
In einer Datenbank sind Cartoon-Videoclips hinterlegt, deren Bilder mit dem Contour-Based SD, sowie dem *Dominant Color Descriptor* [26] beschrieben worden sind. Man hat die Möglichkeit, über ein Webformular innerhalb einer Reihe von Comic-Figuren eine auszuwählen. Alternativ kann der Benutzer die gewünschte Figur in einem Formularfenster selbst zeichnen und colorieren. Anschließend wird mit Hilfe der beiden Deskriptoren die Kontur- und Farbinformation gemäß der jeweiligen Spezifikationen extrahiert und eine Ähnlichkeitsanfrage gegen die in der Datenbank gespeicherten Formen gestartet. Als Ergebnis erhält der Anwender alle Clips, die ähnliche Figuren enthalten, präsentiert und kann sich diese dann herunterladen und ansehen. [20]

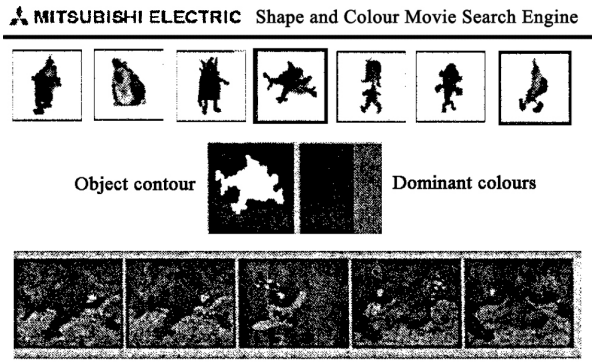

Abbildung 5. Webbasierte Videoclip-Suche mit Hilfe des Contour-Based SD und des MPEG-7 Dominant Color Descriptors. (Nach [19])

5 Zusammenfassung

In dieser Arbeit wurden die vier Deskriptoren zur Beschreibung von Formen, welche im MPEG-7 Visual Standard festgelegt wurden, vorgestellt. Ihre Funktionsweise wurde ausführlich beschrieben und es wurde jeweils ein Beispiel für ein

Ähnlichkeitsmaß gegeben. Des Weiteren wurde gezeigt, dass die vorgestellten Deskriptoren weitgehend dem Kriterium der Invarianz gegenüber den im jeweiligen Kontext üblicherweise auftretenden Verunreinigungen der Daten genügen und damit der Intention ihrer Entwicklung auch in der Praxis entsprechen.

Literatur

1. Hebb, D.O.: The organization of behavior. John Wiley (1949)
2. Koenderink, J., Doorn, A.V.: Dynamic shape. Biological Cybernetics **53** (1986) 383–396
3. Zaharia, T., Prêteux, F.: 3d-shape-based retrieval within the mpeg-7 framework. In: Proc. SPIE Conf. on Nonlinear Image Processing and Pattern Analysis XII. Volume 4304. (2001) 133–145
4. Ullman, S.: High Level Vision. Cambridge, MA: MIT Press (1997)
5. Paquet, E., Murching, A., Naveen, T., Tabatabai, A., Rioux, M.: Description of shape information for 2-d and 3-d objects. Signal Processing: Image Communication **16** (2000) 103–122
6. Vranic, D., Saupe, D.: 3d model retrieval. In: Proc. of Spring Conference on Computer Graphics and its Applications (SCCG2000). (2000) 89–93
7. Kim, W., Kim, Y., Kim, Y.: A new region-based shape descriptor. In: ISO/IEC MPEG99/M5472. (1999)
8. Zaharia, T., Prêteux, F.: 3d versus 2d/3d shape descriptors: A comparative study. In: Dans SPIE Conference on Image Processing : Algorithms and Systems III. Volume 5298. (2004)
9. Dorai, C., Jain, A.: Shape spectrum-based view grouping and matching of 3d free-form objects. IEEE Trans. on PAMI **19** (1997) 1139–1145
10. Koenderink, J.: Solid shape. The MIT Press (1990)
11. Ullmann, J.R.: An algorithm for subgraph isomorphism. Journal of the ACM **1** (1976) 31–42
12. Sclaroff, S.: Deformable prototypes for encoding shape categories in image databases. Pattern Recognition **30** (1997) 627–641
13. Hebert, M., Ikeuchi, K., Delingette, H.: A spherical representation for recognition of free-form surfaces. IEEE Trans. on PAMI **17** (1995) 681–690
14. Zhang, D., Hebert, M.: Harmonic maps and their applications in surface matching. Proeedings of the Int. Conf. on CVPR (1999) 524–530
15. Murao, T.: Descriptors of polyhedral data for 3d-shape similarity search. In: Proposal P177, MPEG-7 Proposal Evaluation Meeting. (1999)
16. Elad, M., Tal, A., Ar, S.: Directed search in a 3d objects database using svm. In: Hewlett-Packard Research Report HPL-2000-20R1. (2000)
17. Zhang, C., Chen, T.: Efficient feature extraction for 2d/3d objects in mesh representation. In: Proc. of the International Conference on Image Processing (ICIP 2001). (2001)
18. Horn, B.: Extended gaussian image. Proc. of the IEEE **72** (1984) 1671–1686
19. Manjunath, B.S., al. In: Introduction to MPEG 7: Multimedia Content Description Language. John Wiley & Sons (2002) 231–260
20. Bober, M.: Mpeg-7 visual shape descriptors. IEEE Transactions on circuits and systems for video technology **11** (2001)
21. Manjunath, B.S., al: 7-8. In: Introduction to MPEG 7: Multimedia Content Description Language. John Wiley & Sons (2002)
22. Bober, M.: Shape descriptor based on curvature scale space. In: MPEG-7 proposal P320. (1998)
23. Mokhtarian, F., Mackworth, A.K.: A theory of multiscale, curvaturebased shape representation for planar curves. IEEE Trans. Pattern Anal. Machine Intell. **14** (1992) 789–805
24. T.Zaharia, Preteux, F., Preda, M.: 3d shape spectrum descriptor. In: ISO/IEC JTC1/SC29/WG11, MPEG99/M5242. (1999)
25. Loop, C.: Smooth subdivision surfaces based on triangles. Master's thesis, Dep. of Mathematics, Univ. of Utah (1987)
26. Sikora, T.: The mpeg-7 visual standard for content description - an overview. IEEE Trans. on Circuits and Systems for Video Technology **11** (2001) 696–702